비지도 대조 학습에서 한국어 문장 표현을 위한 특수 토큰 컷오프 방법의 유효성 분석

한명수[O1], 정유현[1], 채동규

한양대학교 인공지능학과
{myngsoo, robo0725, dongkyu}@hanyang.ac.kr

On the Effectiveness of the Special Token Cutoff Method for Korean Sentence Representation in Unsupervised Contrastive Learning

Myeongsoo Han[1], Yoo Hyun Jeong[1], Dong-Kyu Chae
Dept. of Artificial Intelligence, Hanyang University

요 약

사전학습 언어모델을 개선하여 고품질의 문장 표현(sentence representation)을 도출하기 위한 다양한 대조 학습 방법에 대한 연구가 진행되고 있다. 그러나, 대부분의 대조학습 방법들은 문장 쌍의 관계만을 고려하며, 문장 간의 유사 정도를 파악하는데는 한계가 있어서 근본적인 대조 학습 목표를 저해하였다. 이에 최근 삼중항 손실 (triplet loss) 함수를 도입하여 문장의 상대적 유사성을 파악하여 대조학습의 성능을 개선한 연구들이 제안되었다. 그러나 많은 연구들이 영어를 기반으로한 사전학습 언어모델을 대상으로 하였으며, 한국어 기반의 비지도 대조학습에 대한 삼중항 손실 함수의 실효성 검증 및 분석은 여전히 부족한 실정이다. 본 논문에서는 이러한 방법론이 한국어 비지도 대조학습에서도 유효한지 면밀히 검증하였으며, 다양한 평가 지표를 통해 해당 방법론의 타당성을 확인하였다. 본 논문의 결과가 향후 한국어 문장 표현 연구 발전에 기여하기를 기대한다.

주제어: 문장 표현, 대조 학습, 비지도 학습 방법

1. 서론

비지도 대조학습은 컴퓨터 비전 분야에서의 성공에 이어서 자연어 처리 분야에서도 인상적인 성능을 달성했다 [1, 2]. 대조 학습에서 문장 표현을 학습한다는 것은 단어나 문장과 같은 객체가 의미적으로 유사한 샘플은 서로 가깝게, 그렇지 않은 샘플은 멀리 떨어져 있는 고정 크기의 벡터에 투영해야한다는 목표를 가진다. 이러한 접근은 positive pair를 잘 구성하거나 적절한 대조 목표를 설계하는 것이다. 특히, NT-Xent (normalized temperature-scaled cross entropy) 손실 함수에 기반한 대조 학습은 앞서 설명한 대조학습의 목표에 부합하여 널리 사용되는 손실 함수이다 [3]. 그러나, pair 단위의 구성에만 초점을 맞추는 것은 대조학습의 학습 목적을 충분히 고려하지 못해 문장 간의 상대적 유사성을 모델링할 수 없다는 한계가 있다.

이를 해결하기 위해, 참고문헌 [4]에서는 기존의 대조학습에 삼중 손실 함수를 도입하고 삼중항 (원본 문장, 약한 변형 문장, 강한 변형 문장) 을 구성하기 위한 삭제 기반의 데이터 증강 (augmentation) 방법을 제안했다. 그러나 문장에서 단어를 임의로 삭제하면 문장의 의미가 훼손될 수 있다는 위험성은 여러 차례 제기되었다 [5, 6]. 특히, 참고문헌 [7]에서는 한국어 데이터셋에서 단어 삭제 기반의 데이터 증강의 성능 저하 문제가 보고되었다. 한국어의 경우 영어와는 다르게 교착어로 분류되며 형태적 다양성으로 인한 어휘 표현 때문에 보다 잘 설계된 증강 방식이 요구된다 [8].

비지도 대조학습에서 대표적인 베이스라인으로 통용되는 SimCSE (참고문헌 [9], [그림 2]의 왼쪽) 는 드롭아웃

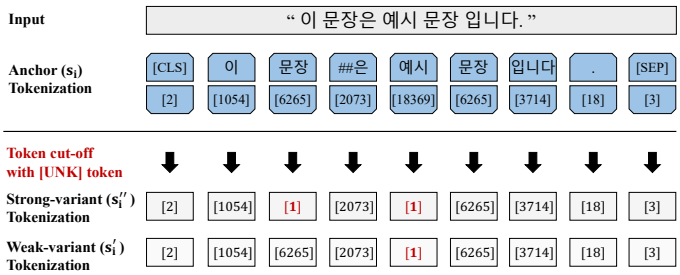


그림 1. 특수 토큰 컷오프 방법 예시. 참고문헌 [10]에 따르면, 강한 변형(strong-variant) 문장은 약한 변형(weak-variant) 문장보다 특수 토큰에 토큰으로 대체되는 비율이 더 많다. 위 예시에서 숫자 [1]은 KLUE-BERT에서 사용되는 [UNK] 특수 토큰을 의미한다.

[1] 공동1저자

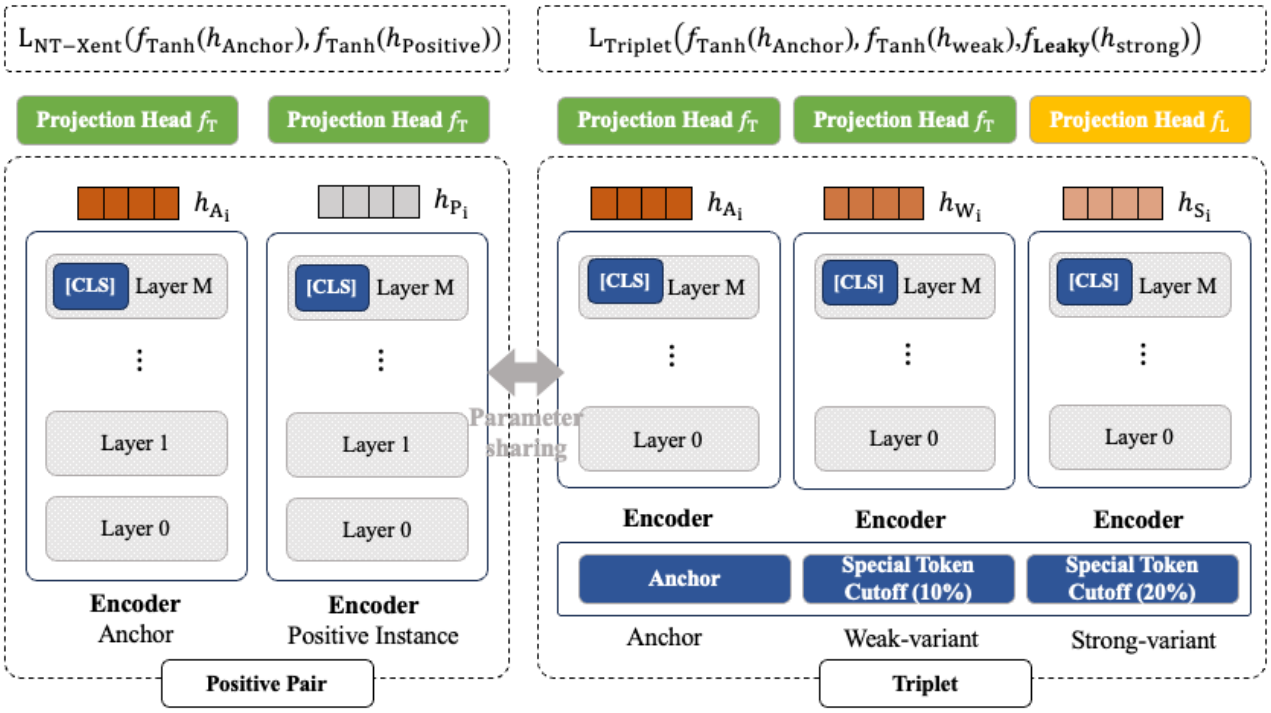


그림 2. SimCSE [9] (왼쪽) 와 본 논문에서 참고한 삼중항 손실 함수 및 토큰 컷오프 기반 데이터 증강 방법 [10] (오른쪽) 의 전체적인 흐름도. 오른쪽 아래 그림은 삼중항 모델링을 위해 특수 토큰 컷오프 방법 (STC, Special Token Cutoff)으로 삼중항을 구성하고, 오른쪽 위 그림은 약한 변형 문장 (weak-variant) 과 강한 변형 문장 (strong-variant) 의 추가적인 섭동(노이즈)을 위해 활성화 함수 변형을 진행한다.

(dropout)을 이용한 간단한 데이터 증강 방식으로도 비지도 대조 학습을 위한 효과적인 positive pair를 생성했다. 참고문헌 [10]은 문장의 상대적인 유사성 고려하는 것 뿐만 아니라 의미론적인 훼손을 최소화 하기 위한 특수 토큰 컷오프 기반의 데이터 증강 및 활성화 함수 변형 방법을 제안하였다. 여기서 컷오프 [28]란 특정 토큰의 값을 0으로 마스킹하는 방식으로, 그림 [1, 2]와 같이 데이터셋에 대한 직접적인 전처리 없이 토큰 단위에서 원본 데이터를 증강할 수 있다. 토큰 컷오프는 대조학습에서도 효과가 검증된 바 있었다 [23].

이에 본 논문은 참고문헌 [10]에서 제안한 특수 토큰 컷오프 방법이 한국어 기반 대조학습에서도 유효한지를 한국어 데이터셋을 통해 확인하고, 다양한 평가 지표를 통해 해당 방법론의 타당성을 검증하고자 한다. 본 논문에서 진행한 실험을 통해 도출한 결론은 다음과 같다. 먼저, 참고문헌 [10]에서 제안한 방법론이 한국어 데이터셋에서도 더 나은 한국어 문장 표현을 만들어 낼 수 있음을 검증하였다. 더하여, 이러한 검증을 기반으로 다음과 같은 분석 결과를 제시한다. 첫째, 다양한 특수 토큰을 통해 삼중항을 구성한 결과 참고문헌 [10]과 마찬가지로 [UNK] 특수 토큰이 한국어 기반 실험에서도 가장 좋은 성능을 보였다. 둘째, 자기 유사도 변화 지표 [24]를 통해 [UNK] 토큰이 대조학습 이후 가장 문맥화 (contextualization) [24] 가 덜 된 토큰임을 확인하였고, 변형 문장을 만들기에 가장 적절하다는 사실을 발견하였다. 셋째, 대조학습 이후 문장 표현 시각화 결과에 따르면 참고문헌 [10]에서 제안된 방법은 한국어 기반 평가에서도 실제로 문장의 상대적 유사성을 잘 구분하여 판별력이 증가하였으며, 새로운 의미 공간을 만들어 내는 현상을 보인다는 점을 확인하였다.

2. 관련 연구

2.1 대조학습

초기 문장 표현 연구에서는 주로 Word2Vec[11]에 기반하여 임베딩 공간(embedding space)에서 가장 가까운 문장을 예측하고, N-gram 임베딩[12]을 통해 의미 정보를 포착하거나 전체 임베딩을 생성하여 문장 표현을 학습했다 [13]. 사전학습 언어모델이 도입되면서 많은 연구들을 통해 문맥화 된 표현(contextualized representation)을 생성하려는 시도가 늘어나고 있으며, 이러한 시도는 전이 학습에서 주목할 만한 성과로 이어졌다 [14, 15]. 일반적으로 이러한 표현은 BERT의 마지막 레이어에 평균 풀링(average pooling)을 적용하거나 [16], [CLS] 특수 토큰을 활용하는 방식이었다 [17]. 그러나 사전학습 언어모델 기반 표현은 주어진 공간에서 이방성(anisotropy) [19] 특성을 지니고 있기 때문에 문장 유사도 [18]와 같은 태스크에서 기존의 정적 임베딩 방법 (e.g., Word2Vec) 보다 낮은 성능을 보인다. 구체적으로, 이방성이란 사전학습 언어 모델을 통해 도출된 표현들이 맥락성(contextuality)을 고려하기 때문에 고차원 원뿔 공간의 형태로 문장 표현이 임베딩 되는 것을 의미한다 [19].

이 문제를 해결하기 위해 몇몇 연구들은 BERT의 문장 표현을 직접 사용하는 대신 후처리(post-processing) 방

법을 통해 문장 표현을 개선하려고 시도했다 [17, 20]. 특히, 대조학습을 기반으로 한 최근 연구들은 고차원 원뿔 공간에서 정렬성(alignment)과 균일성(uniformity)을 개선하여 이방성 문제를 완화하며 문장 유사도 데이터셋에서 주목할만한 결과를 제시하고 있다 [4, 9, 21].

2.2 데이터 증강 기반 비지도 대조학습

대조학습의 목표는 의미가 유사한 두 입력이 표현 공간에서 서로 가깝게 위치하도록 하고, 의미가 유사하지 않은 입력들은 서로 멀리 떨어지도록 학습하는 것이다 [22]. 따라서 대조 학습의 목표를 잘 수행할 수 있는 새로운 손실 함수를 제안하거나, 의미론적으로 유사한 positive pair 및 의미론적으로 유사하지 않은 negative pair를 구성하는 방법이 중요하다. 특히, positive pair를 생성하는데 있어서 비지도 대조 학습에서는 원본 데이터를 증강하여 활용한다. SimCSE [9]는 비지도 대조 학습을 위해 positive pair를 생성하는데 있어서 dropout을 활용한 데이터 증강 방식을 제안하였으며 그 효과성을 실험적으로 확인하였다. 또한 ConSERT [23]는 고품질 문장 표현을 위해 적대적 공격, 토큰 셔플링, 컷오프 등의 증강 방식을 통해 positive pair를 구성한다.

그러나 위와 같은 방법들은 여전히 pair 단위로 증강이 이뤄지기 때문에, 문장의 상대적인 유사성을 반영하는데 한계가 있었다. 이에 ArcCSE [4]는 문장 간의 의미론적 순서를 반영할 수 있도록 문장 내에서 연속되는 단어를 제거하는 방법으로 삼중항 (positive triplet)을 구축한다. 한편, 참고문헌 [10]은 문장 내에서 일부 단어를 제거하는 방식으로 원본 데이터를 증강하는 것의 위험성을 지적하며, 삼중항 구성을 위한 특수 토큰 컷오프 방법 및 활성화 함수 변형 방법을 제안한다.

3. 평가 지표

본 논문에서는 참고문헌 [10]에서 제안한 특수 토큰 컷오프 기법의 성공에 대한 심층적인 분석을 수행하고, 한국어 기반 데이터셋 및 언어모델에서의 효과성을 검증하고자 한다. 이를 위해 우선 아래와 같은 평가 지표들을 사용한다.

3.1 Alignment와 Uniformity

최근 참고문헌 [24]에서는 문장 표현의 질을 측정하기 위해 대조학습과 관련한 두 가지 속성인 정렬성(alignment)과 균일성(uniformity)을 제시했다. 먼저, positive sample의 분포가 주어질 때 전자의 경우는 두 문장의 표현 사이의 기대 거리를 계산하며, positive sample의 거리를 가깝도록 학습시켜야 하는 대조학습의 목표에 부합한다. 한편, 후자는 주어진 데이터의 분포 내에서 문장 표현이 얼마나 균등하게 분포되어 있는지를 나타낸다. 일반적으로 표현이 문맥화될 수록 이방성의 특성이 뚜렷해지는데, 사전 학습 언어 모델을 활용하여 도출된 표현들은 맥락성을 고려하기 때문에 문장 내의 각 단어가 가진 고유한 의미를 보존하지 못할 수 있다. 때문에, 데이터의 분포 내에서 표현이 균등하게 분포되어 있는지에 대한 여부를 파악하는 것은 표현의 질을 판단하는 하나의 수단이 된다. 균일성과 적합성은 아래와 같이 정의된다.

$$\text{Align} \triangleq \mathbb{E}_{(x,\ y)\sim p_{pos}}[\|f(x)-f(y)\|_2^{\alpha}],\ \alpha>0 \quad (1)$$

$$\text{Uniform} \triangleq \log \mathbb{E}_{(x,\ y)\sim p_{pos}}\left[e^{-t\|f(x)-f(y)\|_2^{\alpha}}\right],\ t>0 \quad (2)$$

3.2 자기 유사도 변화 (Self-Similarity Change)

앞서 전술된 바와 같이 표현의 공간이 hypersphere에서 고르게 분포하여 각 단어는 문맥성을 반영함과 동시에 의미론적 고유성을 보존해야 한다. 이 때, 문장 내의 각 단어가 다른 문장에서도 얼마나 유사한지를 측정하는 척도로 자기 유사도(self-similarity)를 사용할 수 있다. 예를 들어, 모델의 특정 레이어에서 "여우" 라는 단어가 문맥화 되지 않았다면, 모든 문장에서 동일한 표현이 될 것이며 SelfSim = 1의 값을 가진다 [19]. 반대로 SelfSim이 낮을수록, "여우" 에 대한 표현이 더 잘 문맥화 되었다는 것을 의미한다.

$$\text{SelfSim}_l(\text{w}) = \frac{1}{n^2-n}\sum_j \sum_{k\neq j} \cos\left(f_l(s_j, i_j), f_l(s_k, i_k)\right) \quad (3)$$

$$\text{ssc} = (\text{SelfSim}_{\text{학습후}} - \text{ani}_{\text{학습후}}) - (\text{SelfSim}_{\text{학습전}} - \text{ani}_{\text{학습전}}) \quad (4)$$

본 논문은 참고문헌 [25]에서 제안한 자기 유사도 변화 (ssc, Self-Similarity Change) (식 4)를 통해 학습 이후의 한국어 언어 모델이 기존의 모델보다 얼마나 특수 토큰의 문맥 정보를 잘 반영하는지 알아보고자 한다. Anistropy baseline(ani)은 서로 다른 문맥에서 임베딩 벡터간의 코사인 유사도 분포를 나타내며, 사전 학습 모델이 학습한 문맥 정보가 얼마나 일관되게 반영되었는지를 나타낸다. 즉, 자기 유사도 변화는 대조 학습 이후의 사전 학습 모델과 학습 이전의 모델 간의 self-similarity와 anisotropy baseline 차이를 계산하여, 대조 학습 이후의 모델이 얼마나 각 토큰의 문맥 정보를 잘 반영하고 있는지를 측정하는 지표이다. 이 때, 자기 유사도 변화가 높은 단어들은 학습 이후 문맥화가 많이 반영된 단어들이며, 그렇지 않은 단어들은 문맥화가 적게 일어났음을 의미한다.

4. 실험 환경

4.1 학습 및 평가 데이터

Kowiki[2]: SimCSE[9]와 동일한 학습 환경을 구축하기 위해, 한국어 위키 데이터를 10^6개 랜덤 추출하며, 비지도 대조 학습을 위한 학습 데이터로 사용한다.

KorSTS[26]: 문장 유사도 작업을 위한 최초의 한국어 데이터셋이며, 문장 표현 학습의 평가 데이터로 활용한다. 해당 데이터셋은 SemEval-STS-B 2012-2017[18] 데이

[2] https://dumps.wikimedia.org/kowiki/

터셋을 한국어로 번역하여 생성하였다.

KLUE-STS[27]: KLUE 벤치마크는 한국어 사전 학습 언어 모델을 평가하기 위한 데이터셋이다. 8가지의 자연어 이해(Natural Language Understanding) 작업으로 구성되어 있으며, 문장 유사도 작업에서 방법론의 성능을 평가하기 위해 KLUE-STS 데이터셋을 사용한다. 테스트 데이터셋은 공개되지 않아, 본 논문은 검증 데이터셋을 평가 데이터셋으로 활용한다.

4.2 실험 세부 사항

실험 진행 시 사용한 배치(batch)의 크기는 256, 학습률(learning rate)는 5e-5, 문장의 최대 길이는 50으로 설정하였다. 또한, 약한 변형 문장과 강한 변형 문장을 생성하기 위해서 특수 토큰 컷오프 비율은 각 20%, 40%로 설정하였다. 이외의 모든 하이퍼파라미터는 SimCSE [9] 논문의 실험 세부 사항과 동일하다.

5. 결과 및 분석

5.1 한국어 데이터셋에서의 특수 토큰 컷오프 기법 유효성 실험

[표 1]은 참고문헌 [10]에서 제안한 특수 토큰 컷오프 방법과 활성화 함수 변형의 유효성을 앞서 설명한 한국어 문장 유사도 데이터셋에 적용하여 실험한 결과이다. 참고문헌 [10]을 따라서 어휘 사전 내 등록되지 않은 단어를 위해 사용되는 [UNK](unknown을 의미) 토큰을 컷오프 실험에 사용했다.

실험 결과 기존의 한국어 기반 베이스라인 성능에 비해 향상된 것을 확인할 수 있었다. 특히, 활성화 함수의 변형 없이 특수 토큰 컷오프 방법만을 활용했을 때에도 해당 방법론의 효과를 확인할 수 있다. 이를 통해 특수 토큰 컷오프 기반의 데이터 증강이 한국어 문장 표현에 있어서도 긍정적인 영향을 주는 것을 알 수 있었다.

Method / Test set	KOR_{STS}	$KLUE_{STS}$	Avg
KLUE $BERT_{base}$	36.54	36.42	36.48
SimCSE-**KLUE $BERT_{base}$**	68.60	69.53	69.06
STC-**KLUE $BERT_{base}$**	**69.54**	70.08	**69.81**
w/o variant-act	69.21	**70.21**	69.71

표 1. 학습 데이터에 따른 특수 토큰 컷오프 (STC) 및 활성화 함수 변형 (variant-act) 방법의 테스트 데이터 성능 결과.

[표 2]는 [UNK] 토큰 외에도 KLUE-BERT에 등록된 다른 특수 토큰을 활용한 실험 결과이다. 실험 결과 미등록 단어를 처리하기 위해 사용되는 [UNK] 토큰이 특수 토큰 가운데 문장의 상대적인 유사성을 고려하는데 가장 적절하게 활용되었음을 보여준다. [SEP] 토큰 또한 좋은 결과를 보였으나 [UNK] 토큰을 사용했을 때의 성능 향상 폭이 더 컸다.

Method	Special Token	STS Task Avg (KOR_{STS}/$KLUE_{STS}$)
SimCSE KLUE $BERT_{base}$	[UNK]	**69.81** (**69.54** / 70.08)
	[SEP]	69.60 (68.92 / **70.28**)
	[PAD]	69.69 (69.27 / 70.11)

표 2. 다양한 특수 토큰을 활용한 실험 결과

5.2 특수 토큰 컷오프 방법의 자기 유사도 변화 측정 결과

[표 3]은 SimCSE와 특수 토큰 컷오프 방법을 사용했을 때 특수 토큰의 자기 유사도 변화를 보여준다. [UNK]토큰의 자기 유사도 변화 값이 양수로 관측된 것과 가장 큰 값을 가진 것을 관찰할 수 있었다. 이는 [UNK]토큰이 문맥화가 덜 되었음을 의미하며, 데이터 증강을 위한 변형 문장을 생성하는데 가장 적합함을 의미한다. 한편, 특수 토큰 컷오프 방법을 사용하면 특수 토큰에 대한 자기 유사도 변화 값이 더욱 상승한 것을 확인할 수 있다. 이러한 결과는 특수 토큰이 대조학습 과정에서 더욱 문맥화가 덜 되었음을 의미하고, 임베딩 공간의 구조를 결정하는데 중요한 역할을 한다고 해석 가능하다 [25]. 이러한 토큰은 문장 임베딩의 주요 맥락을 대표할 수 있기 때문에, 문장의 상대적 유사성 평가에서 핵심적인 역할을 한다. 결론적으로, 입력 토큰을 제거하였던 기존 연구 [4]와 달리, [UNK] 특수 토큰을 활용한 증강 방법은 변형된 임베딩 공간의 구조를 결정하는데 중요한 역할을 하기 때문에 원본 문장과 의미론적으로 더 멀어진 문장을 생성했음을 확인할 수 있다.

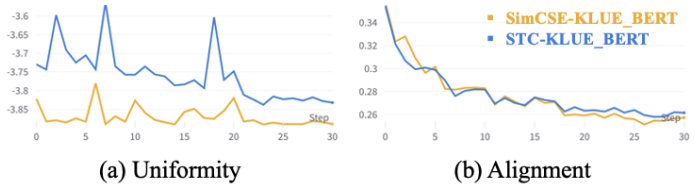
따라서 본 실험 결과들을 통해 [UNK]토큰의 한국어 기반 데이터 증강에서의 유효성을 확인할 수 있었고, 대조학습 과정에서 [UNK] 토큰이 의미 공간의 앵커(anchor) 토큰으로 사용될 수 있음을 시사한다. 또한 비지도 대조학습에서 특수 토큰을 활용한 삼중항의 구성은 한국어 데이터셋에서도 문장 간의 상대적 유사성을 안정적인 학습 대상으로 삼을 수 있다는 가능성을 보인다.

	UNK	SEP	PAD	CLS
SimCSE	0.045	-0.050	-0.037	-0.072
STC-UNK	0.075 **(+ 67%)**	-0.028 **(+ 44%)**	-0.034 (+ 8%)	-0.050 **(+ 30%)**
STC-SEP	0.029 (- 35%)	-0.047 (+ 6%)	-0.023 (+ 37%)	-0.056 (+ 28%)
STC-PAD	0.045 (- 0%)	-0.036 (+ 28%)	-0.020 **(+ 43%)**	-0.056 (+ 28%)

표 3. KoWiki 데이터셋에 대해서 학습한 모델의 훈련 전후 특수 토큰의 자기 유사도 변화 결과

또한 앞서 기술한 바와 같이, 특수 토큰 컷오프 및 활성화 함수 변형 방법은 학습 이후 [UNK] 토큰의 자기 유사도 변화를 가장 많이 발생시켰다 (+ 67%). 특히, [UNK] 토큰을 활용한 방법은 문장 전체를 표현하는 [CLS] 토큰의 변화를 가장 많이 이끌었는데, 이는 [UNK] 토큰의 존재로 인해 모델이 문장의 전반적인 의미를 이해하는데 도움을 받고 있으며, 문장의 문맥 정보를 효과적으로 포착하고 문장 전체를 의미 있게 표현할 수 있다는 신호로 해석할 수 있다.

5.3 BERT-base를 활용한 특수 토큰 컷오프 방법의 Uniformity 및 Alignment 측정 결과



(a) Uniformity (b) Alignment

그림 3. SimCSE 와 STC 방법론의 Uniformity & Alignment 비교

[그림 3]은 SimCSE [9]와 참고문헌 [10] (STC)의 의 학습 과정에서 측정한 균일성(uniformity)과 정렬성(alignment) 결과이다. 문장 표현 측면에서 참고문헌 [10]의 방법은 SimCSE에 비해 균일성이 낮음 (값이 높음을 의미)을 보인다. 균일성이 낮은 이유는 문장 임베딩에 대한 [UNK]토큰의 영향력이 커질수록 [UNK]토큰이 문맥화가 덜 되었기 때문에 문장 임베딩의 핵심 역할로 사용되었을 가능성이 높고, 각 문장 임베딩들이 [UNK] 토큰을 기준으로 한 위치로 수렴했을 것으로 보인다. 이러한 이유로 인해 이방성의 특성을 크게 완화하지는 않은 것으로 추측한다. 이러한 특성에 대한 직관적인 결과는 [그림 4]에 제시되어 있다.

5.4 문장 표현 시각화

[UNK] 토큰은 특수 토큰들 중에서 가장 문맥화가 덜 된 토큰이며, 문맥 내에서 다른 토큰들을 한 위치로 수렴하도록 유도한다. 때문에, [UNK] 토큰을 활용한 약한 변형 문장, 강한 변형 문장 생성은 원본 문장과는 의미론적으로는 비슷하지만 분명한 차이가 존재한다. 이에 본 논문은 특수 토큰을 활용하여 삼중 항(원본 문장, 약한 변형 문장, 강한 변형 문장)을 생성하고, 서로 다른 방법론을 통해 임베딩 공간에서의 차이를 시각화 한다.

시각화 결과는 [그림 4]에서 확인할 수 있다. 특수 토큰을 사용한 컷오프 방법을 대조 학습에 적용할 때, 기존 베이스라인과도 임베딩 공간에서의 차이가 있음을 확인할 수 있었다. [그림 4]의 (b, d, f)와 같은 임베딩 공간이 형성된 것은 [표 3]에 제시된 특수 토큰의 자기 유사도 변화로 인해 각 문장의 임베딩들이 한 위치로 수렴한 것으로 해석된다. 예를 들어, [그림 4]의 (a)의 경우, SimCSE는 약한 변형 문장, 강한 변ㅗ형 문장을 원본 문장과 유사한 위치에 임베딩한 반면, 참고문헌 [10]에서 제안한 방법론은 새로운 의미 공간에 임베딩하는 모습을 보인다.

6. 결론

본 논문은 기존의 영어 문장 표현 학습과 관련된 연구를 한국어 기반 데이터셋 및 언어모델에 적용하여 유효성을 검증하고, 한국어 문장 표현에 미치는 영향을 분석하기 위해 비지도 대조 학습에서 널리 사용되는 평가 지표를 활용했다. 실험 결과에 따르면, 특수 토큰, 그 중에서도 [UNK] 토큰은 대조 학습 과정 중에 문맥화가 덜 발생할 뿐만 아니라 임베딩 공간의 구조를 결정하는데 중요한 역할을 한다. 이러한 분석 결과는 문장의 상대적 유사성 학습 목표에 부합하는 데이터 증강 기법으로 확인되었으며, 이러한 발견은 한국어 문장 표현 연구의 발전에 기여할 수 있을 것으로 기대한다.

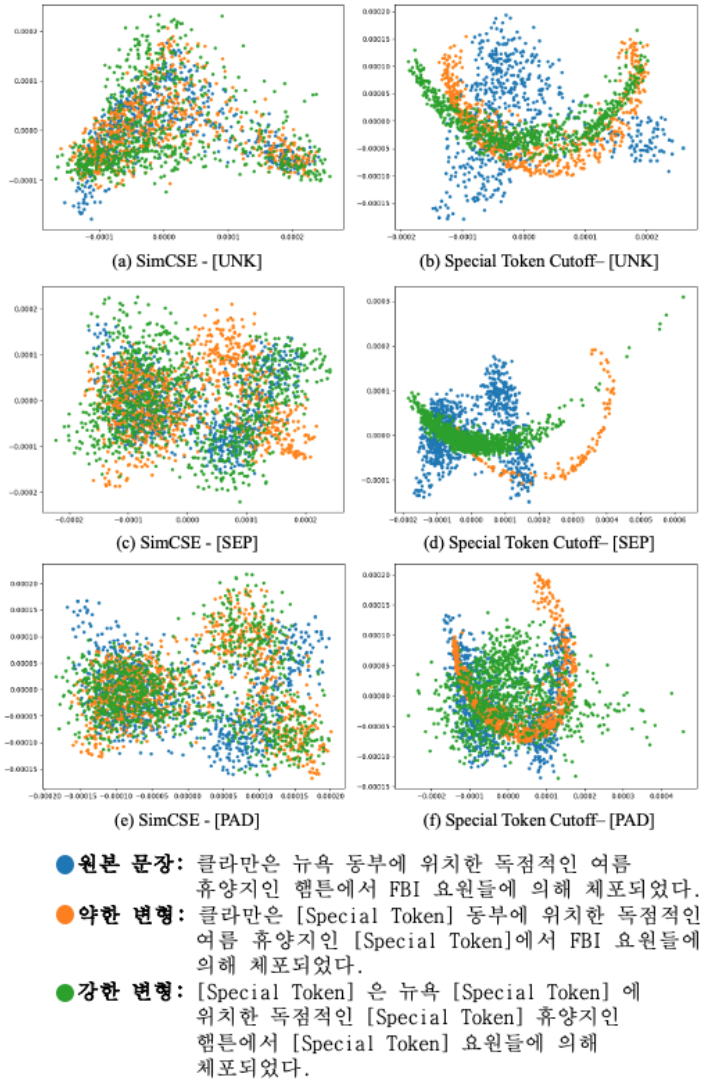


그림 4. 문장 표현 시각화 결과. 서로 다른 3개의 특수 토큰 {UNK, SEP, PAD}에 대해서 {원본 문장, 약한 변형, 강한 변형} 데이터셋을 생성하여, SimCSE와 참고문헌 [10] 의 방법론으로 훈련된 모델을 통해 도출된 문장 표현에 대해서 t-SNE를 통해 시각화하였다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-01373, 인공지능대학원지원(한양대학교))을 받아 수행되었습니다.

참고문헌

[1] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.

[2] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
[3] Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." *arXiv preprint arXiv:1807.03748* (2018).
[4] Zhang, Yuhao, et al. "A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.
[5] Wei, Jason, and Kai Zou. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." *arXiv preprint arXiv:1901.11196* (2019).
[6] Karimi, Akbar, Leonardo Rossi, and Andrea Prati. "AEDA: an easier data augmentation technique for text classification." *arXiv preprint arXiv:2108.13230* (2021).
[7] 서가은, 오하영. “한국어 데이터를 활용한 data augmentation.” *Journal of the Korea Institute of Information & Communication Engineering* 27.4 (2023).
[8] Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." *arXiv preprint arXiv:2008.03979* (2020).
[9] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." *arXiv preprint arXiv:2104.08821* (2021).
[10] 한명수, 정유현, 채동규. “비지도 대조 학습에서 삼중항 손실 함수 도입을 위한 토큰 컷오프 기반 데이터 증강 기법.” 한국정보처리학회 학술대회논문집, 30(1), 618-620.
[11] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[12] Brown, Peter F., et al. "Class-based n-gram models of natural language." *Computational linguistics* 18.4 (1992): 467-480.
[13] Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi. "Unsupervised learning of sentence embeddings using compositional n-gram features." *arXiv preprint arXiv:1703.02507* (2017).
[14] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
[15] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).
[16] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).
[17] Li, Bohan, et al. "On the sentence embeddings from pre-trained language models." *arXiv preprint arXiv:2011.05864* (2020).
[18] Cer, Daniel, et al. "Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation." *arXiv preprint arXiv:1708.00055* (2017).
[19] Ethayarajh, Kawin. "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings." *arXiv preprint arXiv:1909.00512* (2019).
[20] Su, Jianlin, et al. "Whitening sentence representations for better semantics and faster retrieval." *arXiv preprint arXiv:2103.15316* (2021).
[21] Giorgi, John, et al. "Declutr: Deep contrastive learning for unsupervised textual representations." *arXiv preprint arXiv:2006.03659* (2020).
[22] Hadsell, Raia, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping." *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*. Vol. 2. IEEE, 2006.
[23] Yan, Yuanmeng, et al. "Consert: A contrastive framework for self-supervised sentence representation transfer." *arXiv preprint arXiv:2105.11741* (2021).
[24] Wang, Tongzhou, and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere." *International Conference on Machine Learning*. PMLR, 2020.
[25] Xiao, Chenghao, Yang Long, and Noura Al Moubayed. "On Isotropy, Contextualization and Learning Dynamics of Contrastive-based Sentence Representation Learning." *Findings of the Association for Computational Linguistics: ACL 2023*. 2023.
[26] Ham, Jiyeon, et al. "KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding." *arXiv preprint arXiv:2004.03289* (2020).
[27] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." *arXiv preprint arXiv:2105.09680* (2021)
[28] Shen, Dinghan, et al. "A simple but tough-to-beat data augmentation approach for natural language understanding and generation." *arXiv preprint arXiv:2009.13818* (2020).